ABSTRACT

Generation of longer cDNA fragments from SAGE tags for gene identification (GLGI) is disclosed. This method converts SAGE tags, which are about 10 base pairs in length, into their corresponding 3' cDNA fragments covering hundred bases. This added information provides for more accurate genome-wide analysis and overcomes the inherent deficiencies of SAGE. The generation of longer cDNA fragments from isolated and purified protein fragments for gene identification is also disclosed. This method converts a short amino acid sequence into extended versions of the DNA sequences encoding the protein/protein fragment and additional 3' end sequences of the gene encoding the protein. This additional sequence information allows gene identification from purified protein sequences. The invention also provides a high-throughput GLGI procedure for identifying genes corresponding to a set of unidentified SAGE tags.